(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－95796

(43)公開日　平成11年(1999)４月９日

(51)Int.Cl.[6]	識別記号	ＦＩ	
G10L 5/04		G10L 5/04	F
3/00		3/00	H

審査請求　未請求　請求項の数４　ＯＬ　（全７頁）

(21)出願番号　特願平9－250857

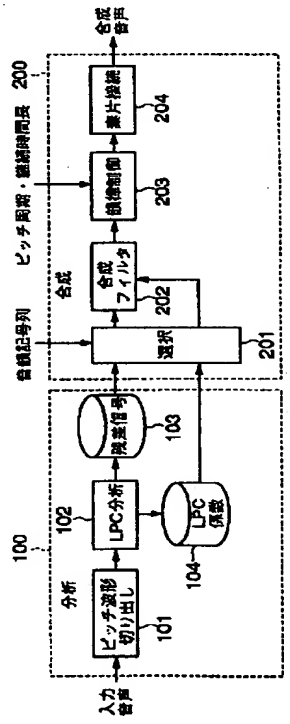
(22)出願日　平成９年(1997)９月16日

特許法第30条第１項適用申請有り　1997年７月18日～７月19日　社団法人情報処理学会主催の「情報処理学会研究報告」において文書をもって発表

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地
(72)発明者　赤嶺　政巳
兵庫県神戸市東灘区本山南町８丁目６番26号　株式会社東芝関西研究所内
(72)発明者　籠嶋　岳彦
兵庫県神戸市東灘区本山南町８丁目６番26号　株式会社東芝関西研究所内
(72)発明者　土谷　勝美
兵庫県神戸市東灘区本山南町８丁目６番26号　株式会社東芝関西研究所内
(74)代理人　弁理士　鈴江　武彦　（外６名）

(54)【発明の名称】音声合成方法

(57)【要約】
【課題】合成音声の音質が優れ、かつ音声素片辞書のサイズがコンパクトで、声質の変更も容易な音声合成方法を提供する。
【解決手段】分析部１００においてピッチ波形切り出し部１０１で切り出した音声素片をＬＰＣ分析部１０２に入力して残差信号とＬＰＣ係数の形で表現して、これらクトルパラメータと残差信号の組を残差信号記憶部１０３とＬＰＣ係数記憶部１０４に音声素片辞書として格納しておき、分析部２００では文解析・韻律制御部から与えられた音韻記号列に従って選択部２０１で残差信号とスペクトルパラメータの組を選択し、選択された残差信号を選択されたスペクトルパラメータに従って構成される合成フィルタ２０２に通すことにより音声素片を作成し、この音声素片に対して韻律制御部２０３でピッチ同期波形重畳法によるピッチ周期の制御と継続時間長の制御を行った後、素片接続部２０４で接続して合成音声を生成する。



続して合成音声信号を生成する。

【００１１】韻律制御に際しては、合成フィルタにより得られる音声素片に対してピッチ同期波形重畳法を適用することによりピッチ周期を制御することが好ましい。韻律制御に際し、さらに音声素片の継続時間長を制御してもよい。

【００１２】このような本発明に基づく音声合成方法によると、従来の残差駆動方式の音声合成法では残差信号のレベルで韻律の制御を行っていたのに対して、音声素片のレベルで韻律の制御を行い、かつ韻律制御後の音声素片を接続するため、波形編集方式と同等の音質の合成音声が得られる。

【００１３】この場合、韻律制御におけるピッチ周期の制御にピッチ同期波形重畳法を用いれば、さらに明瞭で高音質の音声合成が可能となる。また、本発明では音声素片辞書として用意する音声素片を残差信号とＬＰＣ係数のようなスペクトルパラメータの組で表現するため、音声素片辞書のサイズもコンパクトとなる。

【００１４】さらに、このように音声素片をスペクトルパラメータと残差信号の組で表現することによって、スペクトルパラメータの操作により合成音声の声質を用意に変更することが可能である。

【００１５】

【発明の実施の形態】以下、図面を参照して本発明の実施の形態を説明する。図１は、本発明による音声合成方法をテキスト音声合成システムに適用した実施形態を示すブロック図である。この音声合成システムは、大きく分けて分析部１００と合成部２００とからなる。

【００１６】分析部１００は、入力される音声波形からピッチ波形を切り出すピッチ波形切り出し部１０１と、切り出されたパッチ波形のＬＰＣ分析（線形予測分析）を行い、残差信号とスペクトルパラメータであるＬＰＣ係数を抽出するＬＰＣ分析部１０２と、ＬＰＣ分析部１０２により抽出された残差信号とＬＰＣ係数の組を音声素片辞書として格納する残差信号記憶部１０３およびＬＰＣ係数記憶部１０４からなる。

【００１７】一方、合成部２００は図示しない文解析・韻律制御部でテキスト合成に供されるテキストを解析して得られる音韻記号列に従って、分析部１００における残差信号記憶部１０３およびＬＰＣ係数記憶部１０４から、個々の音韻記号に対応する組の残差信号とＬＰＣ係数を選択して取り出す音声素片選択部２０１と、選択されたＬＰＣ係数に従って構成され、選択された残差信号を入力として音声素片を作成する合成フィルタ２０２と、作成された音声素片に対して、文解析・韻律制御部から与えられるピッチ周期および継続時間長の情報に従って韻律の制御を行う韻律制御部２０３と、韻律制御後の音声素片を接続して合成音声を生成する素片接続部２０４からなる。

【００１８】次に、図２に示すフローチャートを用いて、分析部１００の詳細な処理手順を説明する。まず、音声波形を分析部１００に入力する（ステップＳ１１）。この音声波形としては、例えば後述するようにして作成された代表音声素片を用いる。

【００１９】次に、ピッチ波形切り出し部１０１で入力の音声波形にピッチ周期長の窓関数を掛けてピッチ周期分の波形を切り出した後、ＬＰＣ分析部１０２でピッチ同期ＬＰＣ分析を行う（ステップＳ１２～Ｓ１３）。この場合、窓関数により音声波形の離散的なスペクトルが平滑化されるため、基本周波数の影響が低減されたスペクトル包絡を得ることができる。

【００２０】ステップＳ１２でのＬＰＣ分析の結果、音声素片がピッチ周期単位の残差信号とＬＰＣ係数の組で表現される。これらのうち残差信号は残差信号記憶部１０３に、ＬＰＣ係数はＬＰＣ係数記憶部１０４に、それぞれ互いに対応付けられて音声素片辞書として格納される（ステップＳ１４）。

【００２１】次に、図３に示すフローチャートを参照して合成部２００の詳細な処理手順を説明する。音声合成に際しては、図示しない文解析・韻律制御部から音韻記号列とピッチ周期および継続時間長（音韻継続時間長）の情報が与えられる。まず、音韻記号列に従って、音声素片辞書を構成している残差信号記憶部１０３とＬＰＣ係数記憶部１０４から、選択部２０１で個々の音韻記号に対応した残差信号とＬＰＣ係数の組を選択して読み出す（ステップＳ２１）。

【００２２】次に、ステップＳ２１で選択されたＬＰＣ係数によって合成フィルタ２０２を構成し、この合成フィルタ２０２にステップＳ２１で選択された残差信号を入力することにより、音声素片を作成する（ステップＳ２２～Ｓ２３）。

【００２３】次に、ステップＳ２３で作成された音声素片に対して、文解析・韻律制御部から与えられるピッチ周期と継続時間長の情報に従って韻律制御部２０３で韻律制御、つまりピッチ周期の制御と継続時間長の制御を行う。

【００２４】具体的には、ステップＳ２３で作成された音声素片に対して、まず波形編集方式と同様にピッチ同期波形重畳法（ＰＳＯＬＡ）を適用してピッチ周期の制御を行う（ステップＳ２４）。ピッチ同期波形重畳法は、例えば文献（６）「F. Charpentier and M. Stella : “Diphone Synthesis Using an Overlap-add Techniqefor Speech Waveforms concateration”, Proc. ICASSP 86, pp. 2015-2018 (1986)」に記載されている公知の手法であるが、本実施形態ではより高音質の音声合成を可能とするため、以下のようにしてピッチ同期波形重畳法に基づくピッチ周期の制御を行う。

【００２５】一般に、合成音声の音質は有声音の滑らかさに負うところが大きい。そこで、本実施形態ではピッチ周期の変化をより滑らかにするために、与えられたピ

【００３７】学習に当たっては、まず事前準備として音声合成単位の音声素片を音声データベース４０１から大量に切り出し、これらを代表音声素片候補４０２とする。同時に、同様な方法で学習のためのトレーニングデータ４０３を作成する。次に、代表音声素片候補のピッチ周期と継続時間長を分析して（４０４）、トレーニングデータ４０３をターゲットに代表音声素片候補のピッチ周期と継続時間長を分析して変更し（４０５）、音声素片を合成する。このような方法で全ての代表音声素片候補４０２と全てのトレーニングデータの組み合わせについて、音声素片を生成する。

【００３８】次に、生成された音声素片のトレーニングデータに対する歪みを計算で求めて評価し（４０５）、全てのトレーニングデータに対する歪みの総和を最小にする代表音声素片を探索して上述の代表音声素片の候補から選択し（４０６）、これを代表素片とする。

【００３９】この学習法は、合成された音声素片の評価結果を音声素片の学習にフィードバックするという意味で、閉ループ学習と呼ぶ。以下に、この学習法で重要になる歪み尺度と代表音声素片の選択法について、具体的な一例を述べる。

【００４０】（歪み尺度）学習の歪み尺度は、主観評価の結果を良く反映するものである必要がある。また、合成音声のパワーは音声合成システムで制御されることから、代表音声素片はパワーが正規化されたレベルで評価する必要がある。このようなことを考慮して、合成音声素片の歪みを次式で定義する。

【００４１】

【数４】

$$e_{ij} = \sum_n \left(r'_j(n) - s'_{ij}(n) \right)^2 \qquad (6)$$

$$r'_j(n) = r_j(n) \Big/ \left(\sum_k r_j(k)^2 \right)^{1/2} \qquad (7)$$

$$s'_{ij}(n) = s_{ij}(n) \Big/ \left(\sum_k s_{ij}(k)^2 \right)^{1/2} \qquad (8)$$

【００４２】ここで、r_jはトレーニングデータを表し、s_{ij}はr_jを目標にした代表音声素片候補u_iによる合成音声素片を表す。

（代表音声素片の選択）合成単位当たりの代表音声素片数をｎ、代表音声素片候補数をＮとすると、代表音声素片の選択はＮ個の候補からｎ個を選ぶ組み合わせの中から次のコスト関数を最小化する代表音声素片の組を一組探索する問題となる。

【００４３】

【数５】

$$c(i_1, \cdots, i_n) = \frac{1}{M} \sum_{j=1}^{M} \min\left(e_{i_1 j}, \cdots, e_{i_n j}\right) \qquad (9)$$

【００４４】ここで、Ｍはトレーニングデータの数である。式（９）のコスト関数を最小化する代表音声素片の組が求まると、全トレーニングデータを代表音声素片に対応するクラスタにクラスタリングすることができる。

【００４５】図５に、４個の代表音声素片候補から２個の代表音声素片を選択する場合の例を示す。この例では、u_1～u_4の任意の二個の組み合わせの中で、u_2とu_3の組み合わせのコスト関数が最小となる。この結果、u_2とu_3が代表音声素片として選択される。

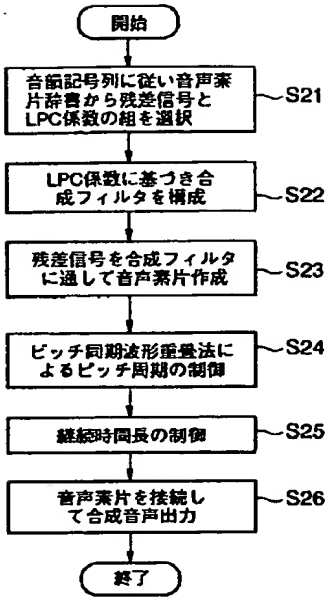
【００４６】（評価実験）ＣＶ、ＶＣのｄｉｐｈｏｎｅを合成単位として、各合成単位に対して上述の方法で１個の代表音声素片を作成する実験を行った。視察により音韻ラベルが付けられた音声データベースからトレーニングに用いる音声素片データと代表音声素片候補を切り出し、前述した閉ループ学習法で計３０２個のＣＶ，ＶＣ代表音声素片を作成した。学習に要した時間はＳｕｎ－Ｕｌｔｒａ２で約１．５時間であった。

【００４７】図６は、合成単位（ＣＶ，ＶＣ）当たりの音声素片数を増加させた場合のコスト関数の値を示しており、この図から音声素片数の増加とともに合成音声の歪みが単調に減少していることが分かる。

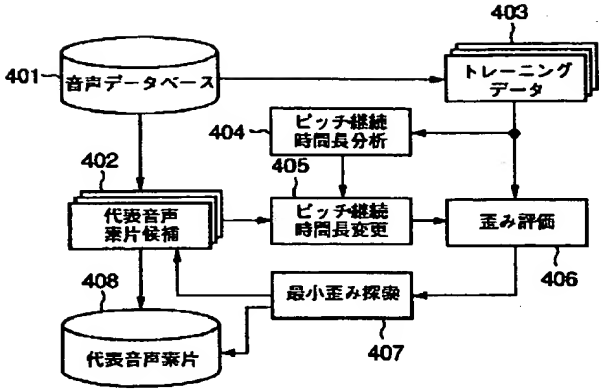
【００４８】従来から、パワーやピッチにより音声素片を使い分けることにより合成音の音質が向上することは知られている。しかし、従来の試行錯誤による方法では、代表音声素片の作成に多大な労力と時間を要し、代表音声素片の数を増やすことは容易ではなかった。

【００４９】これに対し、上述した閉ループ学習法によれば、ラベリングされた音声データが与えられれば短時間で自動的に音声素片の作成ができ、任意の数の代表音声素片を生成することが容易である。しかも、パワーやピッチといった先見的な知識で音声素片の選択を行うのではなく、合成音声の歪みの尺度で選択の規則を作成することが可能である。すなわち、トレーニングデータを選択された代表音声素片のクラスタにクラスタリングし、クラスタ内で共通する要因を抽出することにより音声素片選択の規則を生成することができる。

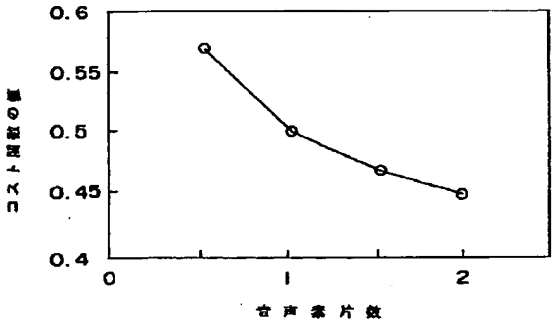
【図３】


開始
音価記号列に従い音声素片辞書から残差信号とLPC係数の組を選択
S21
LPC係数に基づき合成フィルタを構成
S22
残差信号を合成フィルタに通して音声素片作成
S23
ピッチ同期波形重畳法によるピッチ周期の制御
S24
継続時間長の制御
S25
音声素片を接続して合成音声出力
S26
終了

【図４】


403
401
音声データベース
トレーニングデータ
404
ピッチ継続時間長分析
402
405
代表音声素片候補
ピッチ継続時間長変更
歪み評価
406
408
最小歪み探索
代表音声素片
407

【図６】


0.6
0.55
0.5
0.45
0.4
0
1
2
音声素片数

(19) JAPANESE PATENT OFFICE

JPA11-095796

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **11095796 A**

(43) Date of publication of application: **09.04.99**

(51) Int. Cl **G10L 5/04**
G10L 3/00

(21) Application number: **09250857**

(22) Date of filing: **16.09.97**

(71) Applicant: **TOSHIBA CORP**

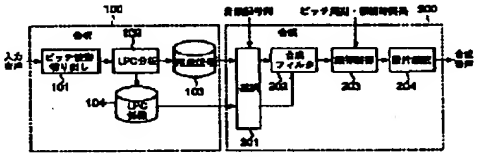
(72) Inventor: **AKAMINE MASAMI**
KAGOSHIMA TAKEHIKO
TSUCHIYA KATSUMI

(54) VOICE SYNTHESIZING METHOD

(57) Abstract:

PROBLEM TO BE SOLVED: To provide a voice synthesizing method by which synthesized voices of superior tone quality can be abtained, the size of a voice element dictionary is compact, and the change in voice quality is easily performed.

SOLUTION: In a analysis section 100, the voice elements segmented by a pitch waveform segmenting section 101 are inputted into an LPC analysis section 102 and expressed in the forms of residual signals and LPC coefficients. A set of these spectrum parameters and the residual signals is stored in a residual signal storage section 103 and an LPC coefficient storage section 104 as a voice element dictionary. In an analysis section 200, a selecting section 210 selects a set of the residual signals and the spectrum parameters in accordance with the phoneme symbol string given by a sentence analysis.rhythm control section. Then, voice elements are generated by passing them through a synthesis filter 202, which is constructed of the selected residual signals and the selected spectrum parameters. Then, a pitch period control by a pitch synchronization waveform superimposing method and a duration length control are conducted in a rhythm control section 203 against the voice elements and synthesized voices are generated by connecting these voice elements an element connecting section 204.



[Title of the Invention]

Voice Synthesis Method

[Abstract]

[Problems]

To provide a voice synthesis method in which the tone quality of synthetic voice is excellent, the size of a voice segment dictionary is compact, and the change of voice quality is easy.

[Solution]

In an analysis part 100, a voice segment cut out in a pitch waveform cutting part 101 is inputted into an LPC analysis part 102 and expressed in terms of a residual signal and an LPC coefficient. A set of spectral parameter and residual signal are stored in a residual signal storage part 103 and an LPC coefficient storage part 104 as a voice segment dictionary. In an analysis part 200, the set of residual signal and spectral parameter is selected in a selection part 201 according to a phonemic symbol string given from a sentence analysis and rhythm control part, the selected residual signal is passed through a synthetic filter 202 conforming to the selected spectral parameter to generate a voice segment, the pitch period and the continuation time length for this voice segment are controlled by a pitch synchronization waveform convolution method in a rhythm control part 203, and

the voice segments are concatenated to generate a synthetic voice.

[Claims]

[Claim 1]

A voice synthesis method comprising expressing a voice segment in terms of a residual signal and a spectral parameter, passing the residual signal through a synthesis filter conforming to the spectral parameter to generate the voice segment, making a rhythm control for the voice segment, and concatenating the voice segments after the rhythm control to generate a synthetic voice.

[Claim 2]

A voice synthesis method comprising expressing a voice segment in terms of a residual signal and a spectral parameter, storing a set of spectral parameter and residual signal as a voice segment dictionary, selecting the set of residual signal and spectral parameter according to a given phonemic symbol string, passing the selected residual signal through a synthesis filter conforming to the selected spectral parameter to generate a voice segment, making a rhythm control for the voice segment, and concatenating the voice segments after the rhythm control to generate a synthetic voice.

[Claim 3]

The voice synthesis method according to claim 1 or 2, wherein the pitch period is controlled by applying a pitch synchronization waveform convolution method to

the voice segment obtained through the synthesis filter in making the rhythm control.

[Claim 4]

The voice synthesis method according to claim 3, wherein the continuation time length of the voice segment is further controlled in making the rhythm control.

[Detailed Description of the Invention]

[0001]

[Field of the Invention]

The present invention relates to a voice synthesis method suitable for the text-speech synthesis. More particularly, the invention relates to a voice synthesis method of generating a synthetic voice from information on the phonemic symbol string, the pitch, and the phoneme continuation time length.

[0002]

[Prior Art]

Generating a voice signal from arbitrary sentence artificially is called a text-speech synthesis. The text-speech synthesis is generally performed at three stages by a language processing part, a phoneme processing part, and a voice synthesis part. The input text, first of all, is subjected to the morphological analysis and the syntactic analysis in the language processing part. Next, the accent and the intonation are processed in the phoneme

processing part, whereby information on the phonemic symbol string, the pitch, and the phoneme continuation time length is outputted. Finally, a synthetic voice is generated from information on the phonemic symbol string, the pitch, and the phoneme continuation time length in the voice signal synthesis part.

[0003]

A voice synthesis method usable for such text-speech synthesis must be a method by which the voice can be synthesized in an arbitrary rhythm for an arbitrary phonemic symbol string. The voice synthesis methods for enabling the arbitrary phonemic symbol string to be synthesized as the voice are divided roughly into an LPC analysis synthesis method and a waveform editing method.

[0004]

The LPC analysis synthesis method involves applying the LPC analysis to a voice signal to acquire an LPC spectral parameter and a residual signal and making the rhythm control and concatenation at the level of the residual signal, as introduced in literature (1) : Ito and Sato, "Pitch control method for voice synthesis using the cut out residual", Sound theory 2-7-18(1989-3), for example. This method has the advantage that the change of voice quality is easy by the operation of the LPC coefficient, and the size

of a voice segment dictionary for the synthesis is comparatively small. However, the tone quality of synthetic voice is poor because the synthetic voice is so-called nasal voice and lacks distinctness.

[0005]

On the other hand, the waveform editing method is the one for synthesizing a voice by changing the pitch period or continuation time length of voice segments cut out from the actual voice waveform and concatenating the voice segments, as introduced in literature (2) : Hirokawa, Hakoda and Sato, "A waveform selection method for waveform editing synthesis in view of spectral continuity", Sound theory 2-6-10 (1990-9), literature (3) : Iwata et al., "Japanese text voice synthesis for personal computer software", Sound theory 2-8-13 (1993-10), and literature (4) : Koyama and Koizumi, "Review on waveform rule synthesis method having a basic unit of VCV", Shingaku technical report, SP96-8 (1996-5), for example. With this method, the sound quality is relatively easy to enhance, and has been vigorously examined.

[0006]

In addition, from the standpoint that the signal processing for analysis and synthesis should not be performed to enhance the sound quality, a method has been offered in which the voice waveforms where the phonemic environment and the rhythm environment are

consistent are concatenated from a database of natural voice in the longest unit (literature (5), N. Campbell and A. W. Black: "CHATR, Arbitrary voice synthesis system of natural voice waveform concatenation type ",Shingaku technical report SP96-7(1996-5)".

[0007]

These methods have the advantage that a synthetic voice of higher sound quality can be produced than the analysis synthesis method, but a problem that the size of the voice segment dictionary is greater. Also, there is a problem that the change of voice quality is difficult because the spectral parameter is not expressed explicitly.

[0008]

This invention has been achieved to solve the abovementioned problems with the prior art, and it is an object of the invention to provide a voice synthesis method in which the tone quality of synthetic voice is excellent, the size of a voice segment dictionary is compact, and the change of voice quality is easy.

[0009]

[Means for Solving the Problems]

In order to achieve the above object, the invention provides a voice synthesis method comprising expressing a voice segment in terms of a residual signal and a spectral parameter such as an LPC coefficient, passing the residual signal through a

synthesis filter conforming to the spectral parameter to generate a voice segment, making a rhythm control for the voice segment, and concatenating the voice segments after the rhythm control to generate a synthetic voice.

[0010]

More specifically, a voice synthesis method comprises expressing a voice segment in terms of a residual signal and a spectral parameter, storing a set of spectral parameter and residual signal as a voice segment dictionary, selecting the set of residual signal and spectral parameter according to a given phonemic symbol string, passing the selected residual signal through a synthesis filter conforming to the selected spectral parameter to generate a voice segment, making a rhythm control for this voice segment, and concatenating the voice segments after the rhythm control to generate a synthetic voice.

[0011]

It is preferred that the pitch period is controlled by applying a pitch synchronization waveform convolution method to the voice segment obtained through the synthesis filter in making the rhythm control. The continuation time length of the voice segment may be further controlled in making the rhythm control.

[0012]

With this voice synthesis method of the invention, the rhythm control is performed at the level of voice segment, and the voice segments after the rhythm control are concatenated, whereby the synthetic voice of the equivalent tone quality to that of the waveform editing method is produced, although the conventional voice synthesis method of residual driving system has the rhythm control at the level of residual signal.

[0013]

In this case, if the pitch synchronization waveform convolution method is employed for the control of pitch period in the rhythm control, the distinct voice synthesis of higher tone quality is enabled. In the invention, since the voice segment prepared as the voice segment dictionary is expressed in terms of the set of residual signal and spectral parameter such as LPC coefficient, the size of the voice segment dictionary is compact.

[0014]

In this way, since the voice segment is expressed in terms of the set of spectral parameter and residual signal, the voice quality of the synthetic voice can be easily changed by the operation of the spectral parameter.

[0015]

[Embodiments of the Invention]

The preferred embodiments of the present invention will be described below with reference to the accompanying drawings. Figure 1 is a block diagram showing a configuration of a text voice synthesis system to which a voice synthesis method according to an embodiment of the invention is applied. This voice synthesis system is roughly composed of an analysis part 100 and a synthesis part 200.

[0016]

The analysis part 100 comprises a pitch waveform cutting part 101 for cutting out a pitch waveform from an input voice waveform, an LPC analysis part 102 for making the LPC analysis (linear prediction analysis) of the cut out patch waveform to extract a residual signal and an LPC coefficient that is a spectral parameter, and a residual signal storage part 103 and an LPC coefficient storage part 104 for storing a set of the residual signal and LPC coefficient extracted by the LPC analysis part 102 as a voice segment dictionary.

[0017]

On the other hand, the synthesis part 200 comprises a voice segment selection part 201 for taking out a set of residual signal and LPC coefficient corresponding to an individual phonemic symbol from the residual signal storage part 103 and the LPC coefficient storage part 104 in the analysis

part 100 according to a phonemic symbol string obtained by analyzing a text used for text synthesis in a sentence analysis/rhythm control part, not shown, a synthesis filter 202 for generating a voice segment by inputting the selected residual signal, the synthesis filter conforming to the selected LPC coefficient, a rhythm control part 203 for making the rhythm control for the generated voice segment according to information on the pitch period and the continuation time length given from the sentence analysis and rhythm control part, and a segment concatenation part 204 for concatenating the voice segments after the rhythm control to generate a synthetic voice.

[0018]

Referring to a flowchart of Figure 2, a detailed processing procedure of the analysis part 100 will be described below. First of all, a voice waveform is inputted into the analysis part 100 (step S11). This voice waveform may be a representative voice segment generated as will be described later.

[0019]

Next, the pitch waveform cutting part 101 cuts out a waveform of pitch period by multiplying the input voice waveform by a window function of pitch period length, and the LPC analysis part 102 makes the pitch synchronization LPC analysis (steps S12 and

S13). In this case, since the discrete spectra of voice waveform are smoothed by the window function, a spectral envelope on which a basic frequency has less influence is obtained.

[0020]

As a result of the LPC analysis at step S12, the voice segment is expressed in terms of the set of residual signal and LPC coefficient in a pitch period unit. The residual signal is stored in the residual signal storage part 103, and the LPC coefficient is stored in the LPC coefficient storage part 104, in which both the residual signal and the LPC coefficient are associated with each other as a voice segment dictionary (step S14).

[0021]

Referring to a flowchart of Figure 3, a detailed processing procedure of the synthesis part 200 will be described below. In the voice synthesis, a phonemic symbol string and information on the pitch period and the continuation time length (phoneme continuation time length) are given from the sentence analysis and rhythm control part, not shown. First of all, the set of residual signal and LPC coefficient corresponding to an individual phonemic symbol is selected and read in the selection part 201 from the residual signal storage part 103 and the LPC coefficient storage part 104 composing the voice

segment dictionary according to the phonemic symbol string (step S21).

[0022]

Next, the synthesis filter 202 is configured conforming to the LPC coefficient selected at step S21, and the residual signal selected at step S21 is passed through the synthesis filter 202 to generate a voice segment (steps S22 and S23).

[0023]

Next, the rhythm control part 203 makes the rhythm control, or the control for the pitch period and continuation time length, for the voice segment generated at step S23 according to information on the pitch period and the continuation time length given from the sentence analysis and rhythm control part.

[0024]

Specifically, first of all, the pitch period is controlled by applying a pitch synchronization waveform overlap-add technique (PSOLA) like a waveform editing method to the voice segment generated at step S23 (step S24). The pitch synchronization waveform overlap-add technique is well known as described in literature (6) : F. Charpentier and M. Stella, "Diphone Synthesis Using an Overlap-add Technique for Speech Waveforms concatenation", Proc. ICASSP 86, pp.2015-2018 (1986), for example. In this embodiment, to allow voice

synthesis of high tone quality, the pitch period is controlled based on the pitch synchronization waveform overlap-add technique in the following way.

[0025]

Generally, the tone quality of synthetic voice greatly depends on the smoothness of voiced sound. Thus, in this embodiment, to make the change of pitch period smoother, the given pitch period is interpolated for a sample unit. Supposing the central time for the j-th frame and the (j+1)-th frame to be t_1, t_2, and the pitch period to be p_1, p_2, when the pitch period is linearly changed, the pitch period p (t) at time t is represented by the following expression.

[0026]

[Equation 1]

$$p(t) = \{(t-t_1)p_2+(t_2-t)p_1\}/(t_2-t_1) \quad (1)$$

Supposing the pitch mark position from t_1 to t_2 to be m_k (k=1, 2, ..,N), the following expression holds.

[0027]

[Equation 2]

$$\int_{m_{k-1}}^{m_k} \frac{2p}{p(t)}\,dt = 2p \quad (2)$$

From the equations (1) and (2), the following expression is obtained.

[0028]

[Equation 3]

$$m_k=m_{k-1}+(m_{k-1},+a)\ (e^b-1) \quad (3)$$

$$a=(t_2p_1-t_1p_2)/(p_2-p_1) \quad (4)$$

$$b=(p_2-p_1)/(t_2-t_1) \quad (5)$$

[0029]

The control of pitch period in the rhythm control part 203 is made by convoluting the voice segments generated through the synthesis filter 202 on the basis of the pitch mark position acquired in this way. That is, the top of voice segment is aligned at each pitch mark position on the time axis, and the voice segments are convoluted with a zero signal. In this case, an overlap portion of adjacent voice segments corresponding to each pitch mark position is added, and the non-overlap portion remains the original voice segment.

[0030]

In the rhythm control part 202, the control of the continuation time length is further made (step S25). In the control of the continuation time length, it is important how the pitch marks of the original voice waveform and the synthetic voice waveform are associated. In this embodiment, the temporal mapping is performed using a function for associating them. With this method, a mapping function is appropriately defined, so that the thinning and interpolation of pitch waveform can be arbitrarily controlled

according to the property of voice segments to be concatenated.

[0031]

Next, the voice segments for which the rhythm control (control of pitch period and continuation time length) is made in the above way are concatenated (step S26). In this embodiment, to reduce a distortion caused by discontinuity of waveform at the concatenated part, the CV and VC segments are employed as the voice segments, whereby the voice segments are concatenated at the vowel steady part. At this time, the pitch waveforms of vowel to be concatenated are added with weight over the entire vowel section. In this way, a synthetic voice in which an arbitrary sentence (text) is translated into the voice signal is produced.

[0032]

Next, a learning method of voice segments according to the invention will be described below. Conventionally, the generation of voice segments relied on a trial and error technique made manually, and it was required that the skilled researcher repeatedly performed a series of operations of cutting out the voice segment from the voice data vocalized in single sound, meaningless word or continuous word over the long time and evaluating the synthetic voice.

[0033]

On the other hand, a method of automatically generating the voice segments from the voice database is a well known phoneme environment clustering (COC: Context Oriented Clustering) method, as disclosed in literature (7) : Nakajima and Hamada, "Rule synthesis method with clustering based on recent sound state", Shingaku theory, D-II, vol. J-72-D-II, No. 8, pp.1177-1179 (1989-8), for example. This method comprises clustering the voice segments cut out from the voice database based on a dispersion of spectral parameter under the restraint condition of phoneme environment and making the centroid of each cluster a representative voice segment.

[0034]

This phoneme environment clustering method has a feature that the representative voice segment can be determined based on a statistical evaluation criterion without relying on the foreknowledge, but does not consider a distortion caused by the control of pitch period that is problematical in the voice synthesis, whereby the tone quality of synthetic voice is not necessarily sufficient.

[0035]

Thus, a learning method of representative voice segment will be described below in which the distortion of synthetic voice is defined, including the distortion caused by making the rhythm control (control of pitch

period and continuation time length) and the distortion is minimized.

[0036]

Figure 4 is a block diagram showing a closed loop learning system for the representative voice segment according to this embodiment. Though this learning method can be practically applied to various synthesizers or synthetic units, an instance where the learning method is applied to learning the CV and VC voice segments for use in the voice synthesis system will be described here. This method comprises obtaining the LPC coefficient and the residual signal for the synthesis filter after generating the voice segment by learning.

[0037]

In learning, first of all, as the preparation, a large amount of voice segments in a voice synthesis unit are cut out from the voice database 401, and made the representative voice segment candidates 402. At the same time, the training data 403 to be learned is generated in the same way. Next, the pitch period and the continuation time length of the representative voice segment candidate are analyzed (404), and the pitch period and the continuation time length of the representative voice segment candidate are analyzed with the training data 403 as target and changed (405) to synthesize the voice segments. In this way, the

voice segments are generated for all the combinations of the representative voice segment candidates 402 and the training data.

[0038]

Next, the distortion of generated voice segment from the training data is calculated and evaluated (405), and the representative voice segment in which the total sum of distortions for all the training data is minimized is searched and selected from among the representative voice segment candidates (406). This selected representative voice segment candidate is made the representative segment.

[0039]

This learning method is called a closed loop learning in the sense of feeding back the evaluation result of synthesized voice segments to the learning of voice segment. In the following, a distortion scale and a selection method of the representative voice segment, which are important in this learning method, will be described using a specific example.

[0040]

(Distortion scale)

The distortion scale of learning is required to reflect the result of subjective evaluation effectively. Since the power of synthetic voice is controlled in a voice synthesis system, it is necessary that the representative voice segment is

evaluated at the level where the power is normalized. In view of this, the distortion of synthetic voice segment is defined by the following expression.

[0041]

[Equation 4]

$$e_{ij} = \sum_{a} (r'_j(n) - s'_{ij}(n))^2 \quad (6)$$

$$r'_j(n) = r_j(n) / (\sum_{k} r_j(k)^2)^{1/2} \quad (7)$$

$$s'_{ij}(n) = s_{ij}(n) / (\sum_{k} s_{ij}(k)^2)^{1/2} \quad (8)$$

[0042]

Where r_j designates the training data, and s_{ij} designates the synthetic voice segment for the representative voice segment candidate u_i with r_j as the target.

(Selection of representative voice segment) Supposing that the number of representative voice segments per synthetic unit is n and the number of representative voice segment candidates is N, the selection of representative voice segment is a problem of searching one set of representative voice segments minimizing the following cost function for all the combinations of choosing n from N candidates.

[0043]

[Equation 5]

$$c(i_1, \ldots, i_n) = \frac{1}{M} \sum_{j=1}^{M} \min(e_{i_1 j}, \ldots, e_{i_n j}) \quad (9)$$

[0044]

Where M is the number of training data. If the set of representative voice segments minimizing the cost function of formula (9) is obtained, all the training data can be clustered into clusters corresponding to the representative voice segments.

[0045]

Figure 5 shows an example of selecting two representative voice segments from four representative voice segment candidates. In this example, the cost function of a combination of u_2 and u_3 is minimized among any two combinations of u_1 to u_4. As a result, u_2 and u_3 are selected as the representative voice segments.

[0046]

(Evaluation experiment)

An experiment was conducted for generating one representative voice segment for each synthetic unit by the above method, with the diphone of CV and VC as the synthetic unit. By inspection, the voice segment data and the representative voice segment candidate used for training were cut out from the voice database with a phoneme label attached, whereby a total of 302 CV and VC representative voice segments were generated by the closed loop learning method. It tool about 1.5 hours to make the learning by Sun-Ultra2.

[0047]

Figure 6 shows the value of the cost function when the number of voice segments per synthetic unit (CV, VC) increases. From this graph, it will be found that the distortion of synthetic voice decreases monotonically as the number of voice segments increases.

[0048]

Conventionally, it is well known that the tone quality of synthetic sound is improved by employing the voice segments according to the power or pitch. However, with the conventional trial and error method, it took a lot of labor and time to generate the representative voice segments, and it was not easy to increase the number of representative voice segments.

[0049]

On the contrary, with the above closed loop learning method, if the labeled voice data is given, the voice segment is automatically generated in a short time, whereby it is easy to generate any number of representative voice segments. And the selection of voice segment is not made according to the prior knowledge such as power or pitch, but a selection rule can be created by the distortion scale of synthetic voice. That is, the selection rule of voice segment can be created by clustering the training data into clusters of selected representative voice segments and extracting a common factor within the cluster.

[0050]

Next, the tone quality of synthetic voice obtained in the above voice synthesis system was evaluated. The generated representative voice segment was inputted as the voice input of Figure 1 to the analysis part, decomposed by the pitch waveform cutting part 101 and the LPC analysis part 102, and stored in terms of the residual signal and the LPC coefficient as the voice segment dictionary in the residual signal storage part 103 and the LPC coefficient storage part 104. When stored, the residual signal and the LPC coefficient were encoded by applying a vector-scalar quantization technique. As a result, the data amount was as small as about 150 kbytes per speaker, which was one-tenth to one-twentieth as compared with the waveform editing method. Accordingly, the voice synthesis system of this embodiment is easily incorporated into a portable information terminal such as PDA or a car navigation system.

[0051]

The subjective evaluation was made at seven stages (-3: very bad to +3: very good) by general subjects, or a total of ten persons (men and women of same number) including seven university students. As a result, the tone quality of synthetic voice obtained by the voice synthesis system of this embodiment is more excellent by 2.5 points on average

in man and woman speakers and various sentences than the voice synthesis system of the conventional cepstrum synthesis method. From the subjects, it is evaluated that the distinctness is greatly improved, and the tone quality is soft and close to natural voice.

[0052]

[Advantages of the Invention]

As described above, with the voice synthesis method of the invention, since the voice segment is expressed in terms of a set of residual signal and spectral parameter such as LPC coefficient, and the rhythm control is made for the voice segment generated by the residual signal and the spectral parameter, the distinct synthetic voice of high tone quality can be generated, the change of voice quality is easily made by the operation of spectral parameter, and the size of the voice segment dictionary is made compact.

[Brief Description of the Drawings]

[Figure 1]

Figure 1 is a block diagram showing a configuration of a voice synthesis system according to one embodiment of the present invention.

[Figure 2]

Figure 2 is a flowchart showing a processing procedure of the analysis side in this embodiment.

[Figure 3]

Figure 3 is a flowchart showing a processing procedure of the synthesis side in this embodiment.

[Figure 4]

Figure 4 is a block diagram for explaining a closed loop learning system of representative voice segment.

[Figure 5]

Figure 5 is a diagram showing a selection example of representative voice segment based on a distortion of synthetic voice segment.

[Figure 6]

Figure 6 is a diagram showing the relationship between the number of representative voice segments and the cost function.

[Description of Symbols]

100 ... voice analysis part

101 ... pitch waveform cutting part

102 ... LPC analysis part

103 ... residual signal storage part

104 ... LPC coefficient storage part

200 ... voice synthesis part

201 ... selection part

202 ... LPC synthesis filter

203 ... Rhythm control part

204 ... voice segment concatenation part

Figure 1

101 Pitch waveform cutting part

102 LPC analysis part

103 Residual signal part

104 LPC coefficient

201 Selection

202 Synthesis filter

203 Rhythm control

204 Concatenation of voice segments

#1 Input voice

#2 Phonemic symbol string

#3 Synthesis

#4 Pitch period and continuation time length

#5 Synthetic voice

#6 Analysis

Figure 2

S11 Input voice.

S12 Cut out pitch waveform.

S13 Pitch synchronization LPC analysis

S14 Store a set of residual signal and LPC coefficient as voice segment dictionary.

#1 Start

#2 End

Figure 3

S21 Select a set of residual signal and LPC coefficient from voice segment dictionary according to phonemic symbol string.

S22 Configure a synthesis filter based on LPC coefficient.

S23 Generate voice segment by passing residual signal through synthesis filter.

S24 Control pitch period by pitch synchronization waveform convolution method.

S25 Control continuation time length.

S26 Output synthetic voice by concatenating voice segments.

#1 Start

#2 End

Figure 4

401 Voice database

402 Representative voice segment candidate

403 Training data

404 Analysis of pitch continuation time length

405 Change of pitch continuation time length

406 Distortion evaluation

407 Minimum distortion search

408 Representative voice segment

Figure 5

#1 Representative segment candidate

#2 Training data

Figure 6

#1 Value of cost function

#2 Number of voice segments

【００５０】次に、上述した音声合成システムで得られた合成音声の音質評価を行った。作成した代表音声素片を図１の音声入力として分析部に与え、ピッチ波形切り出し部１０１およびＬＰＣ分析部１０２を介して残差信号とＬＰＣ係数に分解した形で残差信号記憶部１０３とＬＰＣ係数記憶部１０４に音声素片辞書として蓄積した。蓄積に当たっては、ベクトルースカラ量子化の手法を適用して、残差信号とＬＰＣ係数を符号化した。この結果、データ量は一話者当たり約１５０ｋバイトと、波形編集方式に比べて１／１０～１／２０の非常にコンパクトなものとなっている。従って、本実施形態の音声合成システムはＰＤＡ等の携帯情報端末やカーナビゲーションシステム等へ組み込みことも容易である。

【００５１】大学生７名を含む計１０名（男女同数）の一般の被験者による７段階（－３：非常に悪い～＋３：非常によい）の主観評価の結果、本実施形態の音声合成システムで得られた合成音声の音質は、従来のケプストラム合成方式による音声合成システムに比較して、男女話者及び各種文章の平均で２．５ポイント向上し、明瞭感が大幅に向上するとともに、ソフトでより肉声に近い音質になったとの評価が被験者から得られた。

【００５２】

【発明の効果】以上説明したように、本発明の音声合成方法によれば、音声素片を残差信号とＬＰＣ係数のようなスペクトルパラメータの組で表現し、残差信号とスペクトルパラメータで生成される音声素片に対して音律の制御を行っているため、明瞭で高音質の合成音声を生成できるとともに、スペクトルパラメータの操作により声質の変更が容易であり、さらに音声素片辞書のサイズもコンパクトにすることができる。

【図面の簡単な説明】

【図１】本発明の一実施形態に係る音声合成システムの構成を示すブロック図

【図２】同実施形態における分析側の処理手順を示すフローチャート

【図３】同実施形態における合成側の処理手順を示すフローチャート

【図４】代表音声素片の閉ループ学習システムを説明するためのブロック図

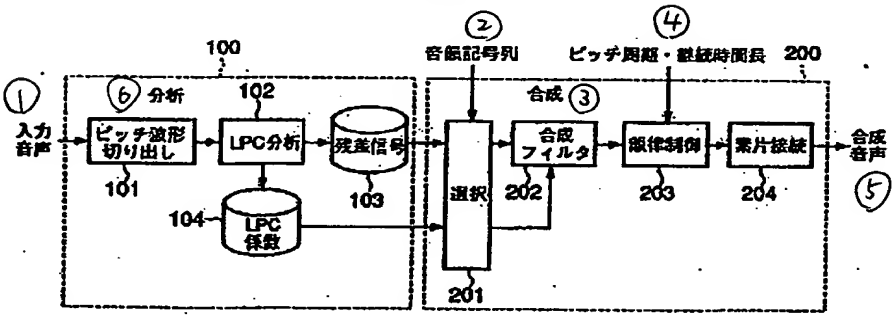
【図５】合成音声素片の歪みに基づく代表音声素片選択の例を示す図

【図６】代表音声素片の素片数とコスト関数の関係を示す図

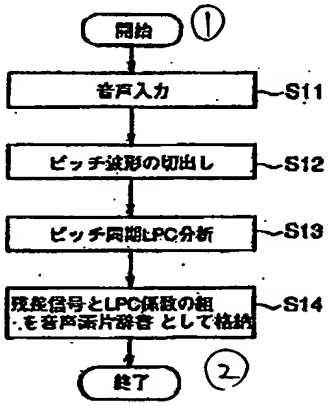
【符号の説明】

１００…音声分析部
１０１…ピッチ波形切出し部
１０２…ＬＰＣ分析部
１０３…残差信号記憶部
１０４…ＬＰＣ係数記憶部
２００…音声合成部
２０１…選択部
２０２…ＬＰＣ合成フィルタ
２０３…韻律制御部
２０４…音声素片接続部

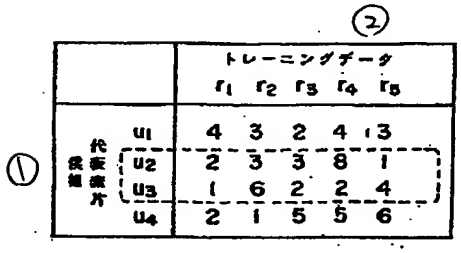
【図１】



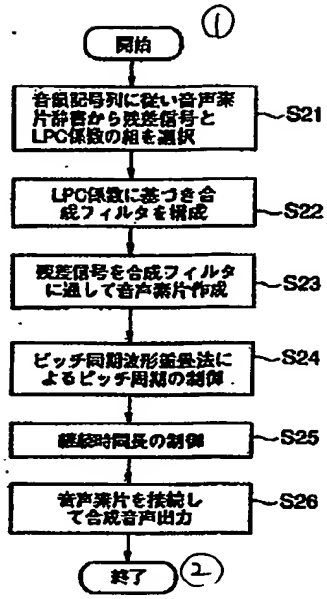
【図２】



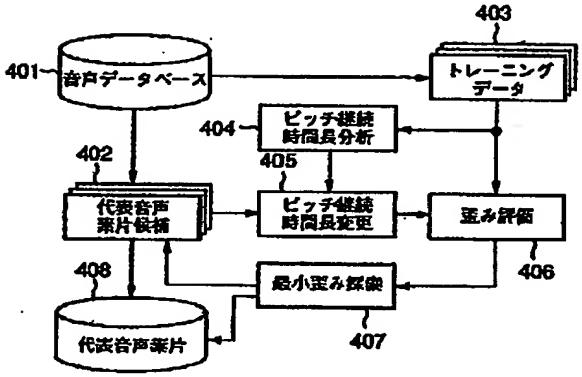
【図５】



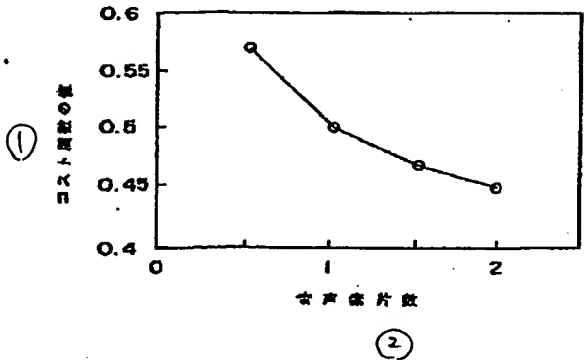
【図３】


①
開始
音韻記号列に従い音声素片辞書から残差信号とLPC係数の組を選択
S21
LPC係数に基づき合成フィルタを構成
S22
残差信号を合成フィルタに通して音声素片作成
S23
ピッチ同期波形重畳法によるピッチ周期の制御
S24
継続時間長の制御
S25
音声素片を接続して合成音声出力
S26
終了
②

【図４】


401
音声データベース
403
トレーニングデータ
404
ピッチ継続時間長分析
405
402
代表音声素片候補
ピッチ継続時間長変更
歪み評価
406
408
最小歪み探索
代表音声素片
407

【図６】


①
0.6
0.55
0.5
0.45
0.4
0
1
2
音声素片数
②